Thanks to the reviewers for their constructive comments.

Reviewer #1. Thanks for the positive feedback. We agree that a multidimensional experiment would be beneficial for the main paper. Note that the paper already has a multidimensional experiment in Appendix E, that will be moved at the end of the experimental section. Moreover we will add some details on the computational cost and fast algorithms to solve the dual problem in Thm. 2 to the end of Sect. 3 (they are now at the end of Appendix B.5).

Reviewer #2. We thank the reviewer for the careful reading and thoughtful comments.

1) As correctly pointed out by Rev2, a model of the form $f_w(x) = (\phi(x)^\top w)^2$ is sufficient for Theorem 3 (and also for Theorem 4). Such formulation would lead to non-convex optimization problems as already noticed by the reviewer. Of course if convexity is not an issue, it is possible to restrict the model to rank-1, keeping all the results of Section 4. However, surprisingly, note that the dual formulation of the problem (Thm. 2) shows that the matrix A also admits a representation in terms of only n degrees of freedom, instead of n^2 (given n coefficients α, the operator A can be recovered via Eq. 11 and then Eq. 7). Thus using the whole matrix A allows on the one hand to still have a representation in terms of n degrees of freedom. On the other hand it leads to a convex problem and allows to control the rank of A via elastic-net regularization. To conclude, we agree that the question raised by Rev2 can be useful to better understand the value of the proposed approach. So we will add all the reasoning above as a remark right after Thm. 2.

2) We would like to point out that we study explicitly the variance of the proposed method in Theorem 5 where we bound the Rademacher complexity of the proposed estimator. Indeed, by using a standard argument based on the Rademacher complexity (see [29] Chapter 26, or [3] paragraph 4.5 and in particular Eq. 13) we can derive the following learning rate. Let the population risk be defined as $R(f) = \mathbb{E}_{x,y}\ell(y, f(x))$ for some G-Lipschitz loss function ℓ and $\widehat{R}_D$ be the empirical version $\widehat{R}_D(f) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i))$ for a given dataset D of n examples. Given a norm $\|\cdot\|_\circ$ (e.g., Frobenius or nuclear), a feature map ϕ and a radius L, define the class of estimators $\mathcal{F}^\circ_{\phi,L} = \{f_A \mid \|A\|_\circ \leq L\}$. Denote by $\widehat{f}_{D,L} = \arg\min_{f\in\mathcal{F}^\circ_{\phi,L}} \widehat{R}_D(f)$ the empirical risk minimization solution over the set $\mathcal{F}^\circ_{\phi,L}$, then

$$\mathbb{E}_D R(\widehat{f}_{D,L}) \leq \inf_{f\in\mathcal{F}^\circ_{\phi,L}} R(f) \;+\; 2\,\mathbb{E}_D\Big[\sup_{f\in\mathcal{F}^\circ_{\phi,L}} |R(f) - \widehat{R}_D(f)|\Big] \leq \inf_{f\in\mathcal{F}^\circ_{\phi,L}} R(f) \;+\; 2G\,\mathcal{R}_n(\mathcal{F}^\circ_{\phi,L}),$$

where $\mathcal{R}_n(\mathcal{F}^\circ_{\phi,L})$ is the Rademacher complexity of the set $\mathcal{F}^\circ_{\phi,L}$ and is bounded by $O(Lc^2/\sqrt{n})$ by Theorem 5 (c is the bounding constant of the kernel, i.e., $c = \sup_{x\in\mathcal{X}}\|\phi(x)\|$). Now, assuming that there exists an operator $A_\star$ with $\|A_\star\|_\circ$ finite (in particular it could be rank-1, i.e. $A_\star = w_\star w_\star^\top$ for some $w_\star$), such that the learning problem is well posed, i.e. $\inf_{f\in C(X)} R(f) = R(f_{A_\star})$, and choosing $L = \|A_\star\|_\circ$, we obtain the learning rate $\mathbb{E}_D R(\widehat{f}_{D,L}) - R(f_{A_\star}) = O(c^2G\,\|A_\star\|_\circ/\sqrt{n})$, that is comparable to the one of kernel linear models [29]. We will add this paragraph as a discussion after Thm. 5, to better clarify the variance and learning rates for the proposed approach.

3) The NCM approach, i.e., approximating a function with non-negative combination of non-negative kernel functions is employed usually in kernel density estimation methods. This model is well known to be quite rigid when the kernel function is non-negative, indeed it can't approximate a density with i.i.d. samples faster than $n^{-1/(d+1)}$ even if the density is arbitrarily smooth (see, e.g., [33]). This happen also when the density is of the form $f_\star(x) = e^{-V(x)}$ with $V(x)$ an infinitely smooth potential. Instead, in this case, according to the Point 2) above, the proposed method has a faster learning rate $O(\|w_\star\|^2/\sqrt{n})$, where ϕ is the Sobolev kernel, since by Thm. 4 there exists a $w_\star$ s. t. $f_\star(x) = (w_\star^\top\phi(x))^2$. We will add this example in Section 4 to clarify the difference between NCM and our method.

4) As suggested by the reviewer we will add a quantitative version of the experiments in the appendix, with 50 i.i.d. repetitions and the resulting error bars. In Appendix E are reported many experimental details. We will add in the main text a more detailed description on how we performed cross validation on both $\sigma \in [10^{-3}, 10^3]$ and $\lambda \in \{0\} \cup [10^{-8}, 1]$ (logarithmic scale, 20 intervals). We would like to note that in the multivariate experiment ($d = 10$), in the appendix, the best parameters found by cross-validation for NCM are $\sigma = 0.5$ and $\lambda = 10^{-4}$. Since the value of σ is not on the left extreme of the range we would exclude that the result obtained by NCM in this experiment was due to oversmoothing. We instead interpret such result in the light of the different learning rates $O(n^{-1/11})$ for NCM and $O(n^{-1/2})$ for our method. In any case we will repeat the experiment (with 50 repetitions) and on a range $\sigma \in [10^{-10}, 10^{10}]$ (log scale with 100 steps). To conclude we would like to recall that we will publish the code on GitHub (python + scipy).

Reviewer #3. We thank the reviewer for the positive feedback. As recalled in the Point 1) and 2) above, the model can be expressed in terms of n degrees of freedom via duality in Thm. 2. In terms of statistical complexity we achieve a learning rate that is similar to the one for linear models (see Point 2). We will add these discussions after Thm. 2 and 5. More details on the computational complexity of the dual formulation (Thm. 2) are at the end of Appendix B.5.

Reviewer #4. Thanks for the thoughtful comments. An extensive explanation about the relation with [4] and its intrinsic limitations is already reported in Appendix F. We will move part of the content in Section 2, where we will also add a short review about main result on SoS programming.